CE888 Assignment 1 (21a2)

November 19, 2021

Set by:	Dr Ana Matran-Fernandez (amatra@essex.ac.uk)
Submission deadline:	Week 21 (see FASER for exact date and time)
Feedback:	Week 24 (3 weeks from submission deadline)
Submission mode:	Electronic via FASER (see Section "Submission" below)

1 Assignment objectives

This document specifies the coursework assignment to be submitted by students taking CE888. The main aims of this assignment are:

1. To think about the project you have been assigned
2. To identify datasets and/or techniques appropriate for your specific problem
3. To undertake data exploration that guides your preprocessing and modelling steps

To do this, you will need to analyse data and present your analysis both through code and in a report.

2 The assignment

Imagine that you are an independent Data Science consultant and you provide scientific advisory and consulting services to companies seeking to apply data analytics to their business activities.

The manager of a company has provided you with a brief description of a project her company is interested in pursuing (the project you have been assigned to). She has asked you to explore the feasibility of their proposed approach and write a report summarising your main findings and recommendations.

You are not expected to finish your full project by Assignment 1. This assignment is designed to get you started on your Data Science project by collecting and preparing your dataset/s for analysis. For this assignment, you should have:

1. Loaded and explored your data, performing appropriate data cleaning and preprocessing steps.
2. Considered your data and the project description to come up with a plan to complete the project.

Your report and code should provide evidence of the two points above.

2.1 The report

The report should be written using an adequate level of English and include the following sections:

1. **Title**: Make sure the title of your report is descriptive of your work. You can be creative with this. Imagine this is going to be read by the company manager. Do not use "Project 1/2/3/Reassessment/Assignment 1/2" as your title.
2. **Executive summary**: Provide a short description of your work, summarising your main findings. A good summary should include a statement of the problem (and its significance), a summary of the methods and results (at this point: which data is going to be used, what type of predictive task is appropriate, how will the model be trained and evaluated), and a short conclusion/recommendation for the company.

   The executive summary should not be longer than 250 words. It should not include references.

3. **Introduction**: Explain the purpose of your work and motivates it. Why is what you are doing important? This section should include references to show that what you are doing is relevant and to back up any claims you make.

   Use a maximum of 600 words for this section (excluding references). Normally there should be no figures or tables in the Introduction.

4. **Data**: Should be divided into subsections, one for each dataset used.

   Describe the dataset/s you are going to use, including how the data was collected (or generated). For each dataset, the following information must be provided (and evidenced through exploration in your code): size of the dataset, types of features, type of problem (e.g., classification, regression, clustering...), whether the data is balanced or not, all the preprocessing steps done to clean and prepare the data for Assignment 2, and what features seem most informative for the problem.

   This is not an exhaustive list and you should give more information that you feel is appropriate. Include figures that show exploratory data analysis and are informative, making sure the axes are labelled and the captions are informative. All figures and tables need to be refered to in the text. If you chose the dataset/s, justify why they are appropriate for the problem you are trying to solve.

   The word limit for this section is 500 words/dataset, excluding references, figures, and tables.

5. **Methodology**: Describe the methodology proposed to the company to solve the problem, providing appropriate justification. Give details of train/validation/test splits (you can do this as percentages or individually for each of the datasets, as appropriate), classifiers/regressors/modelling tools, and details of evaluation methods and metrics that you intend to use. The choice of methods and metrics should be based on the characteristics of your dataset/s. Include a diagram to show the different parts and how the whole project ties together.

   The word limit for this section is 600 words, excluding references, figures, and tables and their captions. Use subsections as needed.

6. **Conclusions**: The conclusions section of the report should include any remarks and recommendations you have for the company about the feasibility of the project proposed. Concentrate on your findings and the relative strengths and weaknesses of the proposed methodology and dataset/s.

   The word limit for this section is 500 words.

2.2 The code

The GitHub repository should include:

- A README file with links/instructions to download the datasets and a description of the repository and how to use/run the code.
- The code that you used to carry out the exploration for your dataset/s. Make sure it is well documented. If you used Jupyter Notebooks, there should be comments on your findings and not just a stream of figures/plots with no justification of why they are informative or what they are showing.

3 Do's and don't's

- This is an individual project and you must work on it by yourself.
- DO read this document twice and check the assignment template and the marking scheme.
- DO read the description of your project at least twice too.
- DO start a thread on the Moodle forum if you have problems/questions about the assignment.
- DO save figures properly from Python (e.g., using `plt.savefig(fname, dpi=1200)`) and include them in the report. I recommend saving them in pdf format for easier formatting on LaTeX.
- DO ensure that each table and figure has an appropriate caption describing it. Refer to them by their number, and not by "figure/table above/below". Tables and figures should be placed at the top or bottom of the page they are in. All tables and figures should be referred to in the main text.
- DO use functions and comments in your code.

- DO write comments and observations if you are using Jupyter Notebooks. If you use python scripts, write comments too!
- DO NOT wait until the last week to get started on the assignment.
- DO NOT copy and paste from other sources (with or without referencing) — this is plagiarism.
- DO NOT copy text from other sources and replace random words — this is also plagiarism (with or without referencing). You must paraphrase the text (and add a reference to it).
- DO NOT include screenshots of your code or code outputs in the report. Any numerical data that you include should be in a suitable graphical or tabular form. You should not include any numerical data that is not relevant to your discussion (do not trivially copy/paste raw output produced by your code).
- DO NOT write your name on the report: use your registration number.

4 A note on paraphrasing

Paraphrasing is more than changing some words in the text (this makes it unreadable and will penalise you). For example, referring to a "random forest" as a "random collection of trees" is not scientific and it does not make sense (and it has been done by students previously!). Other real-life examples include replacing "cross-validation" with "cross-approval", and "deep neural network" with "profound neural organization".

Use your head: read the text, think about the idea you want to convey, and write it down in your own words without looking at the original source. Make sure you add references to the original source/s.

5 Submission

Your work must be submitted to FASER by the deadline given on the system. No other mode of submission is acceptable. Do not wait until close to the deadline to make your submission. Difficulties with the submission system will not be accepted as an excuse for a missing/late submission.

You must submit a report in PDF format (no doc, docx, etc.), adhering to the CE888 Assignment 1 template available on Moodle.

To get the word count per section, you can use: `https://app.uio.no/ifi/texcount/online.php`

6 Marking criteria

This assignment is worth 15% of the module mark and will be assessed according to the marking scheme available on Moodle.